IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 July 2023, accepted 27 July 2023, date of publication 1 August 2023, date of current version 9 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3300895

RESEARCH ARTICLE

CACDU-Net: A Novel DoubleU-Net Based Semantic Segmentation Model for Skin Lesions Detection in Images

SHENGNAN HAO[1], HAOTIAN WU[1], CHENGYUAN DU[1], XINYI ZENG[1], ZHANLIN JI[1,3], (Member, IEEE), XUEJI ZHANG[2], AND IVAN GANCHEV[3,4,5], (Senior Member, IEEE)

[1]Hebei Key Laboratory of Industrial Intelligent Perception, North China University of Science and Technology, Tangshan 063210, China
[2]School of Biomedical Engineering, Shenzhen University Health Science Center, Shenzhen, Guangdong 518060, China
[3]Telecommunications Research Centre (TRC), University of Limerick, Limerick, V94 T9PX Ireland
[4]Department of Computer Systems, University of Plovdiv "Paisii Hilendarski," 4000 Plovdiv, Bulgaria
[5]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, 1040 Sofia, Bulgaria

Corresponding authors: Xueji Zhang (zhangxueji@szu.edu.cn) and Ivan Ganchev (ivan.ganchev@ul.ie)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0135700; in part by the Bulgarian National Science Fund (BNSF) under Grant КП-06-ИП-КИТАЙ/1(KP-06-IP-CHINA/1); and in part by the Telecommunications Research Centre (TRC), University of Limerick, Ireland.

ABSTRACT Skin lesion segmentation is a critical task in the field of dermatology as it can aid in the early detection and diagnosis of skin diseases. Deep learning techniques have shown great potential in achieving accurate lesion segmentation. With the help of these techniques, the lesion segmentation process can be automated, thus reducing the impact of manual operations and subjective judgments. This aids in improving the work efficiency of medical professionals by saving their time and lowering their corresponding effort, and in enabling better allocation of healthcare resources. This paper proposes a novel CACDU-Net model, based on the DoubleU-Net model, for performing skin lesion segmentation better. For this, firstly, the proposed model adopts a pre-trained ConvNeXt-T as an encoding backbone network to provide rich image features. Secondly, specially designed ConvNeXt Attention Convolutional Blocks (CACB) are utilized by CACDU-Net to refine feature extraction by combining ConvNeXt blocks with multiple attention mechanisms. Thirdly, the proposed model utilizes a specially designed Asymmetric Convolutional Atrous Spatial Pyramid Pooling (ACASPP) module between the encoding and decoding parts, using atrous convolutions at different scales to capture contextual information at different levels. The image segmentation performance of the proposed model is evaluated against existing mainstream models on two skin lesion public datasets, ISIC2018 and PH2, as well as on a private dataset. The obtained results demonstrate that CACDU-Net achieves excellent results, especially based on the two core metrics used for the evaluation of image segmentation, namely the Intersection over Union (*IoU*) and Dice similarity coefficient (*DSC*), according to which it surpasses all other models. Moreover, experiments conducted on the PH2 dataset show that CACDU-Net has strong generalization ability.

INDEX TERMS Atrous convolution, attention mechanism, convolutional neural network (CNN), encoding-decoding network, skin lesion segmentation.

I. INTRODUCTION

Skin cancer has three main types: basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanoma, [1]. The etiology of skin cancer is complex, and it often occurs in skin tissues exposed to sunlight. When skin cells lose control of their growth, they can develop into skin cancer, with melanoma being the deadliest type. Young and middle-aged individuals account for about two-thirds of malignant melanoma cases, while individuals aged 65+

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Callico.

account for about one-third. In 2018, the estimated number of melanoma cases was 287,700, with 60,700 deaths, [2]. In recent years, the number of skin cancer incidences has continued to increase. Early diagnosis and timely treatment are the most effective ways to cure melanoma. However, if a person is diagnosed late, the survival rate is only 15% in the advanced stage. Medical researchers have summarized several clinical diagnostic methods for melanoma based on the color, shape, texture, and visual features of pigmented networks and streaks in the skin lesion area under dermoscopy. These methods include the asymmetry, border, color, and differential structure (ABCD) rules [3], pattern analysis [4], the Meng's method [5], and the seven-point feature method [6]. However, the complexity of the skin lesion area, such as body hair, borders, and blood vessels, greatly hinders medical personnel from making accurate judgments. Thus, skin lesion segmentation remains a challenging task.

Currently, skin lesion segmentation methods can be divided into two categories [7], [8]: (i) traditional machine learning (ML) methods [9], such as edge-based [10], region-based [11], threshold-based [12], [13], and clustering-based segmentation methods [14], [15]; and (ii) deep learning (DL) methods. Traditional ML image segmentation methods analyze the differences between the foreground and background of the image and manually design features from information such as grayscale, contrast, and texture in the image for segmentation. With the rise of ML, segmentation methods that extract features purely manually became the mainstream methods at that time. However, these methods can miss out a lot of detailed information. Also, due to some limitations, such as the complexity of designing and extracting the features, ML technology is limited in further development in the field of segmentation. DL can fully utilize the intrinsic information of images and thus gradually became the preferred technology in the field of image segmentation. With the rapid development of convolutional neural networks (CNNs) [16] in the field of image segmentation, there are already specialized medical segmentation models that have achieved great success in on-site and assisted diagnosis. Significant breakthroughs have also been made in the field of skin lesion segmentation. Ghafoorian et al. [17] proposed a multi-branch deep CNN (DCNN) for extracting multi-scale contextual features. However, their network is too shallow to extract high-resolution features. With the development of batch normalization (BN) [18] and residual structure [19], the problems of network degradation and gradient disappearance were solved, by making the network deeper. Yu et al. [20] reported that deep architectures can extract highly discriminative features for skin lesion segmentation, but these networks ignore global features because they focus on local contexts, thus limiting the use of deep architectures in achieving more accurate results.

Recently, attention mechanisms have become popular in DL for extracting global features to enable accurate segmentation. In [21], attention mechanisms were used in combination with the popular U-Net architecture [22] to select discriminative features by weighting different channels for different organs with varying sizes, shapes, and other features. However, the use of a single attention mechanism failed in lesions with complex features.

The motivation of this paper was to develop a skin lesion segmentation model based on DoubleU-Net, which to employ multi-scale feature extraction modules for the extraction of highly discriminative deep features, on one hand, and attention mechanisms for refining the features extracted by the decoder after upsampling, on the other hand. The result of these efforts was a novel CACDU-Net model (https://github.com/1194449282/CACDU-Net), which demonstrated excellent performance in image segmentation experiments conducted on the ISIC2018 [23] and PH2 [24] public datasets, and our own private dataset.

The main contributions of the paper reflect three aspects:

1) The DoubleU-Net [25] network architecture is improved by employing two U-Net networks, named Network1 and Network2, each consisting of encoder and decoder parts. The latest ConvNeXt-T CNN [26], which utilizes a large 7×7 convolution, followed by downsampling, and extracts features in four stages, is employed in the encoding stage of Network1. Different attention mechanisms, combined with standard convolution and ConvNeXt blocks for feature extraction, are applied in both the decoding stage of Network1 and the encoding and decoding stages of Network2.
2) Specially designed ConvNeXt Attention Convolutional Blocks (CACB) are used to provide attention information in both channel and spatial dimensions, focusing on the lesion itself rather than on irrelevant information such as body hair, bubbles, vessels, and measurement scales. Additionally, the use of a stacked U-shaped architecture perfectly combines multi-level features, capturing long-term dependencies in obtaining a global contextual view to help the network achieve accurate segmentation of skin lesions.
3) A newly designed Asymmetric Convolutional Atrous Spatial Pyramid Pooling (ACASPP) module is utilized between the encoding and decoding parts to provide multi-scale semantic information to the network, which is helpful for identifying lesions of different sizes. Asymmetric convolution is employed by ACASPP in conjunction with dilated convolution, whereby different shapes of asymmetric convolution absorb information from different angles and different dilation rates of dilated convolution capture information at various scales.

II. RELATED WORK

A. MEDICAL IMAGE SEGMENTATION

With the development of artificial intelligence, CNNs have gradually been applied to medical image segmentation. Fully

convolutional networks (FCN) [27] were pioneers in image segmentation being able to predict every pixel, with end-to-end, pixel-to-pixel training, thus solving the problem of spatial resolution. In 2015, Ronneberger et al. [22] proposed a new end-to-end semantic segmentation network called U-Net, based on FCN, which is suitable for medical image segmentation. The difference between U-Net and FCN is that U-Net uses convolutional operations with the same number of layers in the upsampling and downsampling stages and connects the downsampling and upsampling layers by skip-connections. Therefore, the features extracted by the downsampling layers can be directly transmitted to the upsampling layers, thus improving the pixel localization and segmentation accuracy of the network. Specifically, U-Net is a U-shaped symmetric encoder-decoder network that uses skip-connections to merge high-level and low-level semantic features [54]. Zhou et al. [28] proposed UNet++, based on the U-Net framework, using a series of nested and dense skip-path connections between the encoder and decoder subnetworks, further reducing the semantic relationship between the encoder and decoder, and achieving better performance in liver segmentation tasks. U-Net++ densely replaces the cropping and concatenation operations in the skip-connections of U-Net with convolutional operations to obtain better feature information and compensate for the information loss caused by sampling [49]. Inspired by U-Net and ResNet [19], DoubleU-Net [25] adds two encoders and decoders to the U-Net model to improve segmentation accuracy. Later, U2-Net [29], named this way because each encoding and decoding layer is nested with U-Net, has shown significant improvements based on some evaluation metrics. Google transplanted Self-Attention (SA) from natural language processing [30] to computer vision and proposed ViT [31] as the backbone. Due to its powerful feature extraction ability, how to combine ViT and its variants with U-Net to obtain better results has been the focus of researchers in recent years [49]. For instance, Swin-Unet [32] combines Swin Transformer [33] with U-Net, and shows better segmentation results. Similarly, SegFormer [34], built on the Transformer architecture, not only demonstrates high performance but it is also very efficient, achieving state-of-the-art results with fewer parameters than other semantic segmentation models. Recently, in 2022, Huang et al. proposed an efficient hierarchical encoder-decoder network called MISSFormer [35], which, due to its unique design components, exhibits improved ability to capture long-range dependencies and local environments.

In the current paper, a novel CACDU-Net model is proposed, based on DoubleU-Net, which demonstrates improved performance in skin lesion segmentation.

B. ASYMMETRIC CONVOLUTION

Asymmetric convolution is a type of convolution operation used in CNNs. Compared to regular convolution, asymmetric convolution has more adjustable parameters and stronger feature extraction capability. In regular convolution, the kernel is usually square or rectangular, with equal width and height, hence it is referred to as symmetric convolution. In contrast, asymmetric convolution allows the width and height of the kernel to be set to different values, enabling the model to better adapt to features of different shapes. EDA-Net [36] is an efficient asymmetric convolutional dense module that decomposes 3×3 convolution into 1×3 and 3×1 convolutions to reduce computational cost. However, its performance degrades in semantic segmentation. To address this issue, Ding et al. [37] proposed a one-dimensional asymmetric convolution to enhance features in the horizontal and vertical directions, and then aggregate the acquired information into a kernel layer to ensure good image recognition performance. Recently, MACU-Net, proposed by Li et al. in [38], applied asymmetric convolution blocks to the field of semantic segmentation, successfully improving the representational power of convolutional layers.

C. ATROUS CONVOLUTION

In the DL field, atrous convolution (also known as dilated convolution) was first proposed in the DeepLab v1 model [38], [39] in 2014 to increase the receptive field and improve the accuracy of image segmentation. Subsequently, more and more DL models began to adopt dilated convolution. In 2015, Szegedy et al. [40] used dilated convolution in the Inception v3 network, which helped to capture a wider range of contextual information and thus improved the performance of image classification and object detection. In 2016, He et al. [19] proposed ResNet, which can better capture detailed information in images through the use of dilated convolution, which helps improve the performance of tasks such as image classification and object detection. In experiments, ResNet showed excellent results on the ImageNet dataset. To fully utilize the image features extracted by deep and shallow networks, a common solution is to fuse multi-scale features [54]. In semantic segmentation, parallel multi-branch structures are usually used to fuse features with different receptive fields. In the DeepLab v2 model proposed by Chen et al. [41] in 2017, an Atrous Spatial Pyramid Pooling (ASPP) module is used as a simple and effective decoder module for clear segmentation. The ASPP module uses dilated convolution with multiple different sampling rates in parallel and fuses them through pooling operations to extract feature information at different scales. This design can effectively capture object features of different sizes and resolutions, thereby improving the performance of the segmentation model. During the feature extraction process, shallow layers contain small receptive fields to represent geometric details, while deep layers contain large receptive fields to represent semantic information [54]. Based on the ASPP module, an Asymmetric Convolutional Atrous Spatial Pyramid Pooling (ACASPP) module is proposed further in this paper for use by the elaborated CACDU-Net model, which uses a different mechanism to adjust kernel sizes.

D. ATTENTION

Attention is a commonly used technique in DL. It assigns different weights to input information based on their relevance, which can be adjusted under different circumstances. Therefore, attention mechanism has high advantages in scalability and robustness. In the field of medical image segmentation, Oktay et al. [21] proposed Attention-UNet based on the U-Net network, which is a novel attention gate (AG) network for medical image processing that can focus more accurately on the regions of interest, suppress irrelevant features, and highlight useful features. Jie Hu et al. [42] proposed the Squeeze and Excitation (SE) attention mechanism and verified it in multiple computer vision tasks, demonstrating that it can significantly improve the performance and generalization ability of CNN models. Woo et al. [43] proposed the Convolutional Block Attention Module (CBAM). Given an intermediate feature map, CBAM sequentially deduces two independent channel and spatial dimensions, and then multiplies the attention map with the input feature map pixel-wise for adaptive feature refinement. Recently, Transformer models have also been widely applied in the field of medical image segmentation, whose self-attention module captures long-range dependencies, while convolution only collects information from neighboring pixels. However, Transformer requires a large amount of training on large-scale datasets to obtain satisfactory results, which poses difficulties in its application to small medical image datasets. In summary, embedding appropriate attention modules at suitable locations in the network for skin lesion segmentation can reduce the impact of irrelevant information such as body hair and bubbles, and obtain more accurate segmentation results [54].

III. PROPOSED CACDU-NET MODEL

This section first introduces the overall structure of the proposed CACDU-Net model, shown in Figure 1, and then describes the details of each module.

A. OVERALL STRUCTURE

As shown in Figure 1, the proposed CACDU-Net model consists of two stacked U-Net structures, namely Network1 and Network2, which utilize different encodings to extract features and perform skip connections. More specifically, Network1 is used to extract coarser features, while Network2 is utilized to extract finer features. This design allows the model to achieve superior segmentation performance at different scales, thereby improving the overall segmentation accuracy. It is worth noting that the prediction results of Network1 are passed through a *Sigmoid* function to become the weights of Network2's input. Specifically, the *Sigmoid* function is applied to the output of Network1, an image of size 256 × 256×1, to transform it into a weight object of the same size, with values ranging from 0 to 1. Then, a matrix multiplication on this weight object and the input image of size 256 × 256×3 is performed, resulting in the input for Network2, which also has a size of 256 × 256×3. This

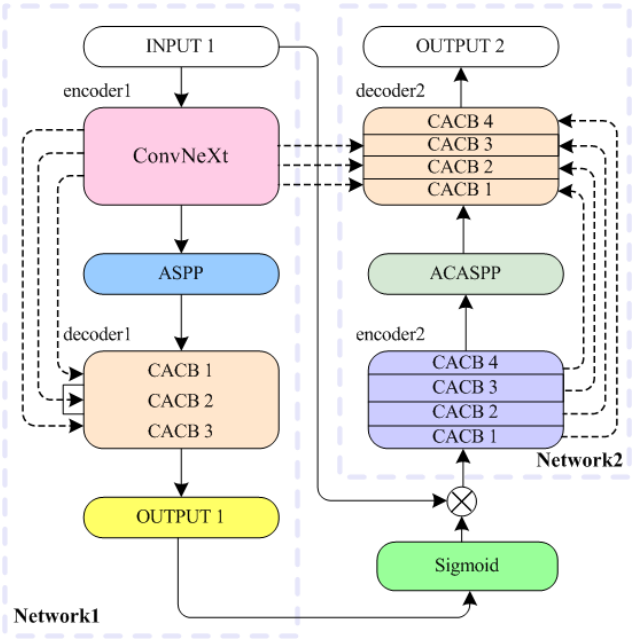


FIGURE 1. The CACDU-net model structure.

enables Network2 to obtain high evaluation scores at an early stage and accelerate the prediction process. The superiority of this architecture is confirmed by the conducted ablation experiments, presented in Section IV.

B. NETWORK1

Figure 2 illustrates the overall structure of Network1. It can be seen that this network adopts a U-shaped architecture composed of encoding, middle, and decoding parts. Skip connections are inserted between the encoding and decoding parts to pass data through. ConvNeXt-T [26], pre-trained on the ImageNet dataset, is used as an encoding part. The middle part performs multi-scale feature extraction using ASPP, with dilation rates set to 6, 12, and 18. Unlike the decoding part of U-Net, all traditional 3 × 3 convolution kernels are replaced by ConvNeXt Attention Convolutional Blocks (CACB), described in Subsection III-E.

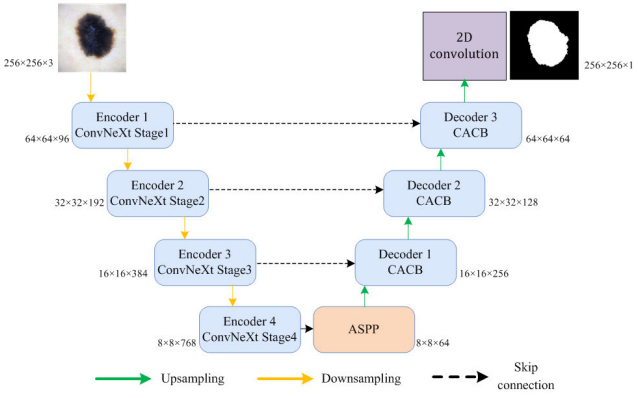


FIGURE 2. The network1 structure.

C. NETWORK2

Figure 3 shows the overall structure of Network2. It can be seen that this network also adopts a U-shaped architecture, which is completely symmetrical and composed of encoding, middle, and decoding parts. Skip connections are used to pass data through, and the network receives, and aggregates features encoded by Network1. The middle part performs multi-scale feature extraction using an ACASPP module, with dilation rates set to 6, 12, and 18. Both the encoding and decoding parts use CACB blocks, which accelerate feature propagation and information flow. Unlike the purpose of Network1, Network2 is designed to further extract features from the input data.

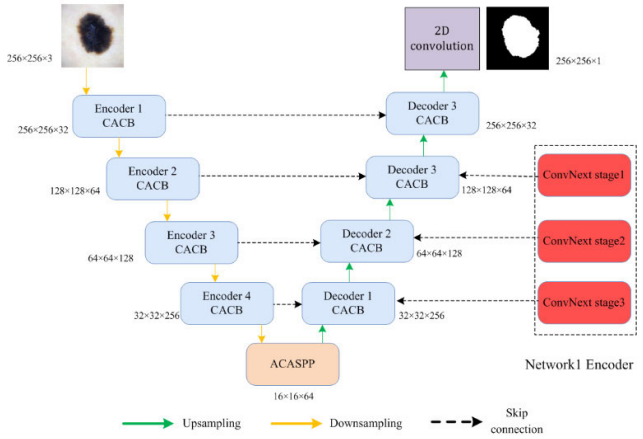


FIGURE 3. The network2 structure.

D. CONVNEXT

ConvNeXt [26] is a CNN, designed to improve feature extraction capability and model performance. Similarly to ResNet, it borrows many successful ideas from the Transformer, but with improved accuracy and efficiency due to the larger kernel sizes and deeper convolutions used. Five versions of the ConvNeXt network were proposed by its authors, namely T/S/B/L/XL, each involving four stages. The only difference between these versions relates to the number of channels and the number of repeated stacked blocks used in each stage [49]. The ConvNeXt-T network refers to the version with the smallest depth and width. Each feature resolution stage of the ConvNeXt-T network consists of multiple residual ConvNeXt blocks (Figure 4a).

As shown in Figure 4b, each ConvNeXt block includes a 7×7 depthwise convolution, two 1×1 layers, and a non-linear Gaussian error linear unit (GELU) activation [44]. Layer normalization (LN) [45] is used before the Conv 1×1 layer. Unlike traditional convolutions, ConvNeXt replaces 3×3 convolutions with 3×3 depth convolutions, uses a reverse bottleneck structure, and employs GELU and LN instead of the Rectified Linear Unit (ReLU) and BN [18], with fewer activation functions and larger convolution kernels up to 7×7. As shown in Figure 4c, the ConvNeXt network

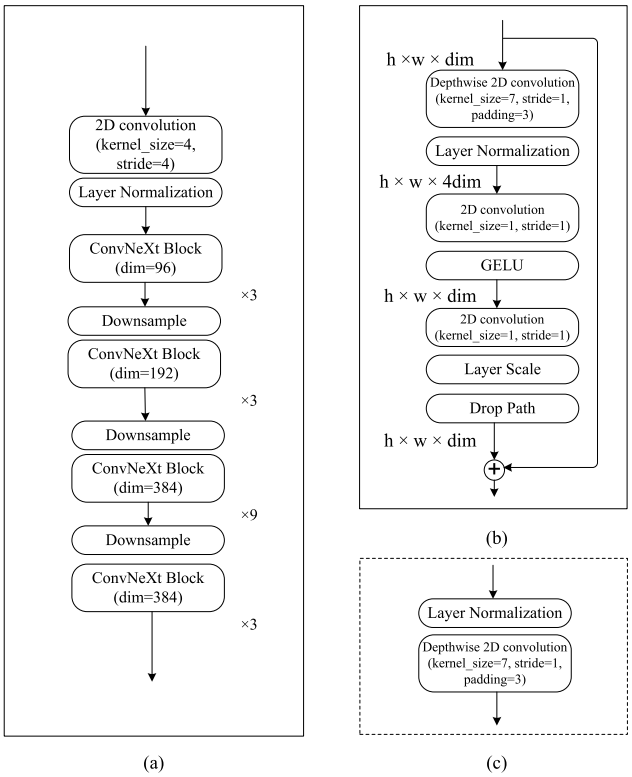


FIGURE 4. (a) The ConvNeXt-T structure; (b) The ConvNeXt block; (c) The ConvNeXt downsample.

utilizes a separate downsampling layer to downsample the features.

As the design of ConvNeXt-T ensures both accuracy and efficiency, it was chosen as the backbone network of the proposed CACDU-Net model.

E. CONVNEXT ATTENTION CONVOLUTIONAL BLOCK (CACB)

Single attention mechanism is insufficient to achieve satisfactory results in complex lesion segmentation. CBAM [43] integrates spatial and channel attention mechanisms to better extract useful feature information, reduce sensitivity to noise and irrelevant features, and improve model accuracy and robustness. Inspired by CBAM, a ConvNeXt Attention Convolution Block, abbreviated as CACB, is proposed here, as shown in Figure 5, which consists of a 3×3 convolution, a ConvNeXt block, and a CBAM channel attention module and a CBAM spatial attention module to extract channel and spatial attention features, respectively. After the 3×3 convolution, BN and ReLU activation functions are performed.

The channel attention block aims to make the neural network focus on global features and suppress unnecessary features such as body hair, measurement scales, blood vessels, and bubbles [54]. This module performs global max pooling and global average pooling on each channel of the input feature map, and then generates two vectors of shape $R^{C\times1\times1}$ (where C represents the number of channels). These two

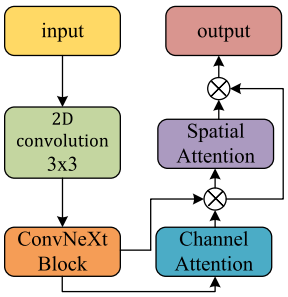


FIGURE 5. The proposed CACB structure.

vectors are then inputted into a multilayer perceptron (MLP), which reduces the number of parameters by sharing weights. The MLP contains only one hidden layer, and its weight vector has a shape of $R^{C/r\times1\times1}$ (where r represents the reduction ratio, which is set to 16 in this paper). The MLP is implemented through two fully connected layers to generate two processed channel attention vectors. Finally, these two vectors are pixel-wise added and processed by a *Sigmoid* activation function, and the feature map size is restored to the same size as the input feature map. The channel attention block functioning is summarized in [54] as follows:

$$\begin{aligned} \mathrm{M_c(F)} &= \sigma\left(MLP\left(AvgPool\left(\mathrm{F}\right)\right) + MLP\left(MaxPool\left(\mathrm{F}\right)\right)\right) \\ &= \sigma\left(\mathrm{W}_1\left(\mathrm{W}_0\left(\mathrm{F}^c_{avg}\right)\right) + \mathrm{W}_1\left(\mathrm{W}_0\left(\mathrm{F}^c_{\max}\right)\right)\right), \end{aligned} \quad (1)$$

where F denotes the input feature map, σ denotes the *Sigmoid* activation function, F^c_{avg} and F^c_{max} denote the feature maps obtained after global average pooling and global max pooling along the channel dimension, respectively, and $W_0 \in R^{C/r\times C}$ and $W_1 \in R^{C\times C/r}$ denote the weights of the MLP.

Different from the channel attention block, the spatial attention block can capture long-term dependency relationships to obtain a global contextual view, and selectively aggregate contextual information according to the spatial attention map to achieve more accurate segmentation performance of skin lesion boundaries [54]. The spatial attention block is more sensitive to lesion edges with similar skin colors in the surrounding area and thus can effectively extract the curve structure features of the edges. More specifically, average pooling and max pooling operations are first performed along the channel axis of the feature map to identify the regions with the maximum information in the feature map. Then, the results of the pooling operations are concatenated to create an efficient feature descriptor. Next, convolutional layers work on the concatenated feature descriptors to generate the spatial attention map, which indicates the positions that should be emphasized or suppressed in the feature map. The specific operations are shown below:

$$\begin{aligned} \mathrm{M_s(F)} &= \sigma\left(f^{7\times7}\left(\left[AvgPool\left(\mathrm{F}\right); MaxPool\left(\mathrm{F}\right)\right]\right)\right) \\ &= \sigma\left(f^{7\times7}\left(\left[\mathrm{F}^c_{avg}; \mathrm{F}^c_{\max}\right]\right)\right), \end{aligned} \quad (2)$$

where $f^{7\times7}$ denotes a convolution operation with a filter size of 7×7, while the channel information of the 2D feature map is represented by $F^c_{\mathrm{avg}} \in R^{1\times H\times W}$ and $F^c_{\max} \in R^{1\times H\times W}$ (where H denotes the height and W denotes the width), respectively.

F. ASYMMETRIC CONVOLUTIONAL ATROUS SPATIAL PYRAMID POOLING (ACASPP)

As reported in [37], square convolution kernels capture features with uneven scales. Specifically, the weights at the center crossing position (i.e., kernel skeleton) have larger magnitudes, while the contributions of the points in the corners to feature extraction are lesser. The design deficiency of the square convolution kernel can be compensated by the use of asymmetric convolution kernels. Shown in Figure 6a, ASPP uses different atrous rates for convolution operations at different sampling rates to extract features within different receptive field ranges, thus capturing multi-scale information. Based on this, the idea of asymmetric convolution proposed in [37] is combined with dilated convolution to design a novel ACASPP module, used by the proposed model to capture features from different receptive fields. As shown in Figure 6b, each dilation rate has two corresponding branches, i.e., 1×3 convolution (horizontal kernel) and 3×1 convolution (vertical kernel), respectively using BN and ReLU to improve numerical stability, in order to obtain a cross-shaped receptive field. The 3×3 convolution in ASPP captures features with a larger receptive field, while the horizontal and vertical kernels ensure the saliency of features on the skeleton, expanding the width of the network [49]. Then, each branch is concatenated, and a 3×3 convolution is used to restore the channel number when recovering the input. Finally, the result is pixel-wise added with the ASPP result and outputted. If $y\,[i]$ denotes the output signal and $x\,[i]$ denotes the input signal, then the atrous convolution can be represented as:

$$y\,[i] = \sum\nolimits_{k=1}^{K} x\,[i + d\cdot k]\cdot w\,[k], \quad (3)$$

where k denotes the kernel size and d denotes the dilation. If $H_{k\times k,r}\,(x)$, where r denotes the dilation rate, represents an operation consisting of Conv2d convolution, BN, and ReLU activation function, then ASPP can be expressed as follows:

$$y_{ASPP} = H_{3\times3,1}\left(\begin{bmatrix} H_{1\times1,1}\,(x)\,; H_{3\times3,6}\,(x)\,; \\ H_{3\times3,12}\,(x)\,; H_{3\times3,18}\,(x)\,; \\ UpSample\left(H_{1\times1,1}\,(AvgPool\,(x))\right) \end{bmatrix}\right) \quad (4)$$

and ACASPP is given by:

$$\begin{aligned} y_{ACASPP} = H_{3\times3,1}&\left(\begin{bmatrix} H_{1\times3,6}\,(x)\,; H_{3\times1,6}\,(x)\,; \\ H_{1\times3,12}\,(x)\,; H_{3\times1,12}\,(x)\,; \\ H_{1\times3,18}\,(x)\,; H_{3\times1,18}\,(x) \end{bmatrix}\right) \\ &+ y_{ASPP}. \end{aligned} \quad (5)$$

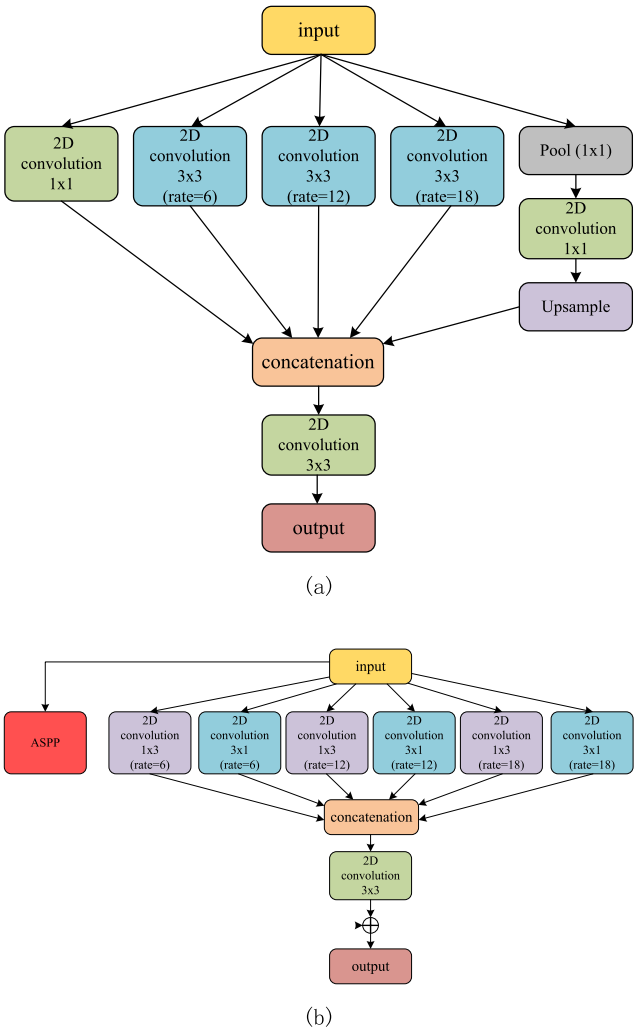


FIGURE 6. **(a) The ASPP module structure; (b) The ACASPP module structure.**

G. LOSS FUNCTION

Composite loss functions, especially those related to dice, often achieve better segmentation results and higher model performance than single loss functions. In medical image segmentation, class imbalance often occurs during experiments, which can result in model training being biased towards densely distributed pixel classes, making it difficult for the model to learn the features of small objects and thus reducing the network's performance. Therefore, a combination of loss functions is used in the conducted experiments for segmentation supervision.

The Binary Cross Entropy (BCE) loss function is widely used in various fields, including semantic segmentation. When using BCE, each pixel is evaluated in sequence, ignoring the contextual labels, and weighting the segmented pixels and background pixels, which greatly helps the network convergence. Because the BCE loss can more effectively calculate the gradient values corresponding to different categories during backpropagation, the problem of gradient disappearance can be better tackled when using it. The BCE loss is defined as follows:

$$L_{BCE} = -\sum_i \left(g_i ln\left(p_i\right) + \left(1 - g_i\right) ln\left(1 - p_i\right)\right), \quad (6)$$

where g_i denotes the segmentation result of pixel i produced by a physician, and p_i denotes the segmentation result of pixel i produced by the network.

The DSC loss function is named after the Dice similarity coefficient (DSC), which is a metric used to evaluate the similarity between two samples. The DSC loss function performs well in scenarios where there is a severe imbalance between positive and negative samples. During model training, it focuses more on the foreground region mining, making the predicted results closer to the actual results. However, if the predicted results in the experimental process are not exactly identical to the true results marked by pixels, there is a possibility of negative impact of the DSC loss function on backpropagation, which makes training a model very difficult. However, using the DSC loss function can reduce the occurrence of overfitting. The DSC loss is defined as follows:

$$L_{DSC} = 1 - 2\frac{\sum_i g_i p_i}{\sum_i g_i + \sum_i p_i}. \quad (7)$$

To accelerate convergence of the network, alleviate the impact of gradient vanishing, minimize the class imbalance issues during backpropagation, and improve skin disease segmentation, a combination of these two loss functions is used for training the proposed model, as follows:

$$L_{=}\frac{1}{2}L_{BCE} + L_{DSC}. \quad (8)$$

IV. EXPERIMENS AND RESULTS

A. DATASETS AND DATA PREPROCESSING

In the experiments, the International Skin Imaging Collaboration Challenge dataset (ISIC2018) [23], the PH2 dataset [24], and a private dataset are used. ISIC2018 is currently the largest skin lesion image dataset in the world, providing professionally annotated digital skin lesion images to facilitate the development of CAD for melanoma and other skin cancers [54]. The PH2 dataset was jointly collected by the Pedro Hispano Hospital in Matosinhos, Portugal, and the Dermatological Services Department of the University of Porto. The private dataset was provided by Peking Union Medical College Hospital, which includes skin lesion images of acne and lupus erythematosus.

ISIC2018 contains 2594 skin microscopy images with segmentation mask labels. For the experiments, this dataset was randomly divided into training, validation, and test sets at a ratio of 7:1:2. Prior to model training, 1/3 of the training set's images was randomly selected to simulate additional random body hair on them by means of a computer program. Additionally, during the training process, operations such as horizontal flipping, vertical flipping, random brightness, Gaussian blur, mean smoothing filtering, and random hue saturation were applied on the ISIC2018 training set (Figure 7). It should be noted that neither of these additional operations were applied on the validation and test sets. The PH2 dataset,

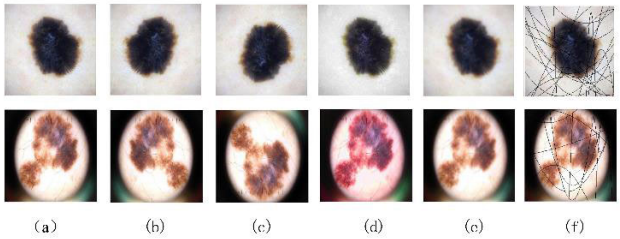


FIGURE 7. Image preprocessing operations: (a) original image; (b) horizontal flip; (c) vertical flip; (d) tone saturation; (e) Gaussian blur; (f) random body hair.

containing only 200 images, served as an additional set for testing the models trained on the ISIC2018 dataset. The private dataset, containing 1010 images, was randomly divided into training, validation, and test sets at a ratio of 8:1:1 for conducting experiments on it. Table 1 shows details of this splitting of the datasets for conducting the experiments for model performance comparison. The ablation study experiments, presented in Subsection IV-D4, were performed only on the ISIC2018 dataset.

TABLE 1. Splitting of datasets into training, validation, and test sets.

Dataset	Training Set's Images	Validation Set's Images	Testing Set's Images	Total Images
ISIC2018 [23]	1816	259	519	2594
PH2 [24]			200	200
Private dataset	808	101	101	1010

B. EXPERIMENTAL ENVIRONMENT

The experiments were conducted in Pytorch version 1.12.1 [46], using Python version 3.10.6, and operating system Ubuntu 22.04. All experiments were conducted on a computer equipped with a 12th Gen Intel®Core™i5-12400 CPU, 16GB RAM, and an NVIDIA GeForce RTX 3060 with 12GB memory. The number of training epochs was set to 150. The Adam optimizer [47] was used with an initial learning rate of 1e-4, weight decay of 1e-6, momentum of 0.9, and batch size of 8. As for the input image size, this was set to 256 × 256 pixels for all models, except for Swin-Unet and MISSFormer for which a size of 224 × 224 pixels was used.

C. EVALUATION METRICS

In the experiments, six evaluation metrics were used to measure the segmentation performance of compared models, namely the Intersection over Union (*IoU*), *DSC*, *accuracy*, *sensitivity*, *specificity*, and *precision*.

IoU, also known as the Jaccard index, is one of the most commonly used metrics in semantic segmentation. IoU is defined as the ratio of the overlap area between the predicted segmentation and the ground truth and their union area. In our case, it is calculated as:

$$IoU = \frac{TP}{TP + FP + FN}, \tag{9}$$

where *TP* (true positives) represents the number of correctly identified pixels as being part of an object (i.e., a skin lesion, in our case), FN (false negatives) represents the number of incorrectly identified pixels as being not part of an object, and *FP* (false positives) represents the number of incorrectly identified pixels as being part of an object.

DSC has become the most universally used metric in the evaluation of image segmentation models. It is defined as twice the overlap area between the predicted segmentation and the ground truth divided by the sum of pixels in both of them. *DSC* is calculated as follows:

$$DSC = \frac{2TP}{2TP + FP + FN}. \tag{10}$$

Accuracy (Acc) is used to evaluate the overall pixel-level segmentation performance, calculated as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{11}$$

where *TN* (true negative) represents the number of correctly identified pixels as being not part of an object.

Sensitivity (Sen) represents the proportion of skin lesion pixels that are correctly segmented, as follows:

$$Sen = \frac{TP}{TP + FN}. \tag{12}$$

Specificity (Spe) is defined as the proportion of non-lesion pixels that are correctly segmented, as follows:

$$Spe = \frac{TN}{TN + FP}. \tag{13}$$

Precision (Pre) represents the proportion of predicted positive samples, as follows:

$$Pre = \frac{TP}{TP + FP}. \tag{14}$$

D. RESULTS AND ANALYSIS

The proposed CACDU-Net model was compared to the mainstream medical image segmentation models by conducting experiments on the aforementioned three datasets, the results of which are displayed in this subsection.

1) ISIC2018 DATASET

The ISIC2018 public dataset contains a relatively large number of skin images, including many difficult-to-segment images [54]. Therefore, the results obtained on this dataset are the most convincing among the three datasets used in the experiments. Thus, the ablation study experiments, presented further below, were conducted only on this dataset.

Table 2 presents the segmentation performance comparison results obtained by state-of-the-art models on this dataset using experimental configurations identical to those used for the proposed CACDU-Net model (the best result on each metric is shown in **bold**). Here, CACDU-Net achieved excellent

TABLE 2. ISIC2018 segmentation performance comparison, based on results obtained by conducted experiments.

Model	*IoU*	*DSC*	*Acc*	*Sen*	*Spe*	*Pre*
U-Net	0.7887	0.8784	0.9508	0.8760	0.9717	0.9182
U-Net++	0.7952	0.8824	0.9528	0.8732	0.9744	0.8914
Attention-UNet	0.7967	0.8833	0.9533	0.8591	0.9790	0.9190
DoubleU-Net	0.8238	0.9014	0.9572	**0.9271**	0.9643	0.8812
U2-Net	0.8221	0.9000	0.9595	0.8817	0.9795	0.9265
Swin-Unet	0.8081	0.8926	0.9560	0.8710	0.9787	0.9182
SegFormer	0.8107	0.8924	0.9571	0.8746	0.9792	0.9198
MISSFormer	0.8175	0.8969	0.9581	0.8742	**0.9812**	**0.9276**
CACDU-Net	**0.8427**	**0.9134**	**0.9641**	0.9065	0.9790	0.9252

results, especially based on the two core evaluation metrics used in image segmentation, namely *IoU* and *DSC*, according to which it outperforms all other models. More specifically, the first runner-up (DoubleU-Net) respectively scored 0.0189 points less for *IoU* and 0.0120 points less for DSC. In addition, based on *accuracy*, the proposed CACDU-Net model also outperformed all models in skin lesion segmentation by leaving the first runner-up (U2-Net) behind by 0.0046 points. According to the other three evaluation metrics used, the proposed CACDU-Net model also performed well in this group, by taking correspondingly the second place on *sensitivity*, third place on *precision*, and fourth (shared) place on *specificity*.

Figure 8 illustrates the loss variation curves of the proposed CACDU-Net model on both the training and validation sets, as well as its DSC and IoU training and validation curves.

Figure 9 shows the Receiver Operating Characteristic (ROC) curves of the compared models, along with their Area Under the ROC curve (AUC) values, achieved on this dataset. As can be seen from this figure, the proposed CACDU-Net model clearly outperforms all other models, as its ROC curve is the closest one to the upper left corner, which indicates the highest overall accuracy.

A visual comparison of the skin lesion segmentation results, achieved by different models on this dataset, is shown in Figure 10.

Table 3 presents the segmentation performance comparison results obtained on the same dataset by other state-of-the-art models, whose results are taken from the specified literature sources (the best result on each metric is shown in **bold**). In this group, CACDU-Net also demonstrated excellent results, especially based on the two core evaluation metrics used in image segmentation (i.e., *IoU* and *DSC*) according to which it outperformed all considered models. More specifically, the first runners-up (ICL-Net and TransCeption, respectively) scored 0.0037 points less for *IoU* and 0.0010 points less for DSC. In addition, based on *specificity* and *precision*, the proposed CACDU-Net model also outperformed all considered models by leaving the first runners-up (TransCeption and M-CSAFN) behind by 0.0046 and 0.0081 points, respectively. Regarding the other

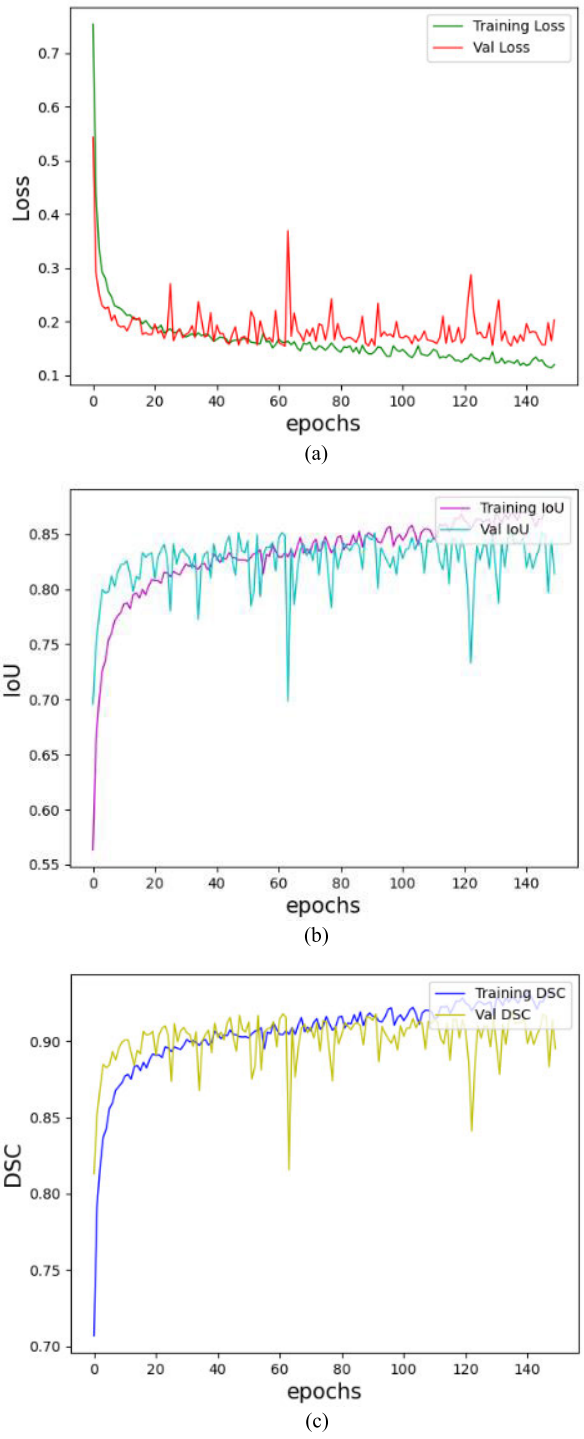


FIGURE 8. Training and validation process of CACDU-Net on the ISIC2018 dataset: (a) training and validation loss curves; (b) IoU training and validation curves; (c) DSC training and validation curves.

metrics, CACDU-Net also performed well, by taking second place on *accuracy* and fourth place on *sensitivity*.

2) PH2 DATASET

In order to test the segmentation performance of the trained model on a new dataset and verify its generalization and

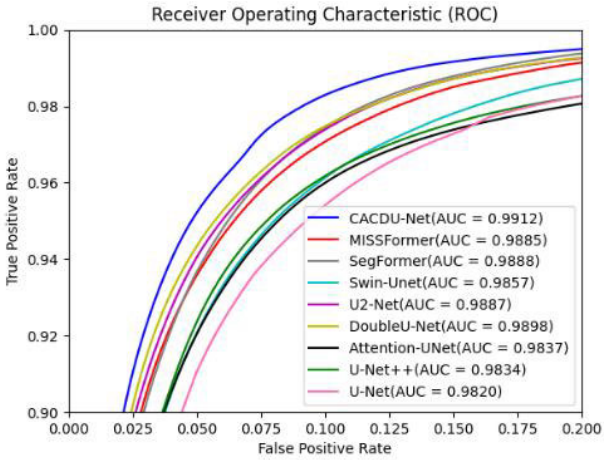


FIGURE 9. ROC curves and AUC values of different models on the ISIC2018 dataset.

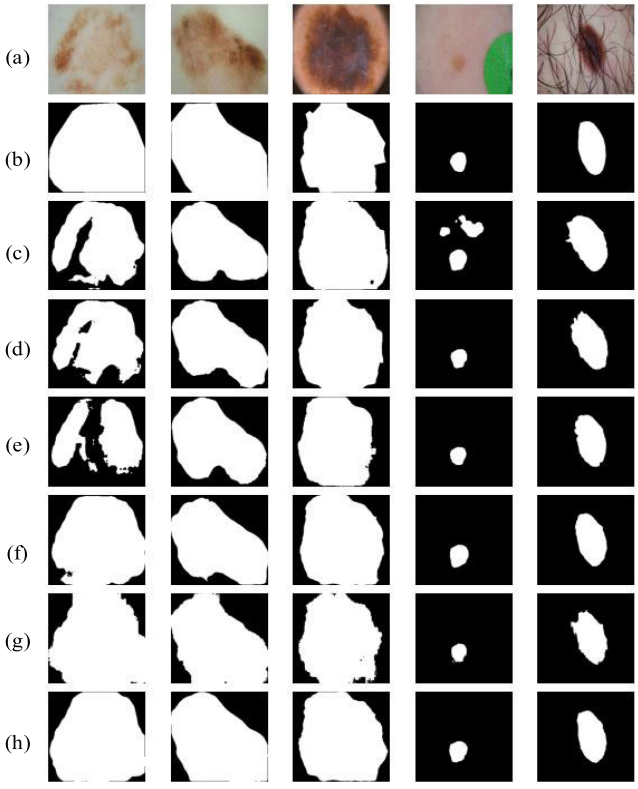


FIGURE 10. Sample segmentations performed on the ISIC2018 dataset: (a) original image; (b) ground truth; (c) U-Net results; (d) U-Net++ results; (e) Attention-UNet results; (f) U2-Net results; (g) Swin-Unet results; (h) CACDU-Net results.

robustness capabilities, experiments were conducted on the PH2 public dataset, which contains only 200 images. For this, the proposed model was trained on the ISIC2018 training set and tested on all PH2 images.

Table 4 presents the segmentation performance comparison results obtained by these experiments (the best result on each metric is shown in **bold**). Again, the proposed CACDU-Net model outperformed all mainstream models

TABLE 3. ISIC2018 segmentation performance comparison, based on results obtained from literature.

Model	*IoU*	*DSC*	*Acc*	*Sen*	*Spe*	*Pre*
Improved U-Net++ [48]	0.8250	0.8680	-	0.8400	-	0.9030
MALUNet [49]	0.8025	0.8904	0.9462	0.8974	0.9619	-
Ms RED [50]	0.8345	0.8999	0.9619	0.9049	-	0.9147
ICL-Net [51]	0.8390	0.9030	0.9440	**0.9410**	0.9290	-
TransCeption [52]	-	0.9124	0.9628	0.9192	0.9744	-
M-CSAFN [53]	0.8364	0.9031	**0.9645**	0.9130	0.9726	0.9171
CACDU-Net	**0.8427**	**0.9134**	0.9641	0.9065	**0.9790**	**0.9252**

TABLE 4. PH2 segmentation performance comparison,based on results obtained by conducted experiments.

Model	*IoU*	*DSC*	*Acc*	*Sen*	*Spe*	*Pre*
U-Net	0.8062	0.8916	0.9276	0.9224	0.9347	0.8674
U-Net++	0.7929	0.8831	0.9238	0.8909	0.9446	0.8814
Attention-UNet	0.7458	0.8505	0.9090	0.8241	**0.9534**	0.8897
DoubleU-Net	0.8109	0.8949	0.9275	**0.9736**	0.9049	0.8300
U2-Net	0.8427	0.9127	0.9405	0.9383	0.9468	0.8941
Swin-Unet	0.8334	0.9084	0.9396	0.9274	0.9482	0.8925
SegFormer	0.8537	0.9197	0.9461	0.9384	0.9533	**0.9052**
MISSFormer	0.8193	0.8989	0.9332	0.9074	0.9514	0.8966
CACDU-Net	**0.8652**	**0.9271**	**0.9504**	0.9706	0.9409	0.8888

based on the two main evaluation metrics, by scoring 0.0115 and 0.0074 points higher than the first runner-up (SegFormer) for *IoU* and *DSC*, respectively. In addition, based on *accuracy*, CACDU-Net also outperformed all mainstream models by leaving the first runner-up (SegFormer) behind by 0.0043 points. According to the other three evaluation metrics used, the proposed CACDU-Net model also performed relatively well, by taking correspondingly the second place on *sensitivity*, sixth place on *precision*, and seventh place on *specificity*. In particular, CACDU-Net performed better in segmenting larger lesions, where U-Net usually failed to segment the entire lesion area, with shapes significantly different from ground truth images. These results demonstrate that the additionally introduced modules indeed improve the segmentation performance and lead to good generalization capabilities.

Figure 11 shows the ROC curves of the compared models, along with their AUC values, achieved on this dataset. As can be seen from this figure, the proposed CACDU-Net model clearly outperforms all other models, as its ROC curve is the closest one to the upper left corner, which indicates the highest overall accuracy.

A visual comparison of the skin lesion segmentation results, achieved by different models on this dataset, is shown in Figure 12.

3) PRIVATE DATASET

Next, experiments were conducted on the private dataset. Compared with the ISIC2018 dataset, this dataset contains

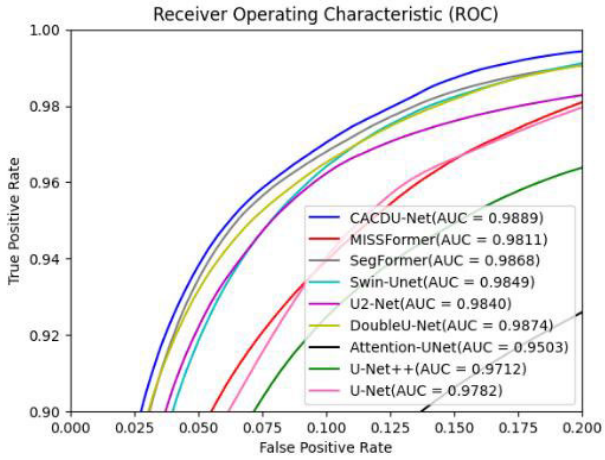


FIGURE 11. ROC curves and AUC values of different models on the PH2 dataset.

TABLE 5. Private dataset segmentation performance comparison,based on results obtained by conducted experiments.

Model	*IoU*	*DSC*	*Acc*	*Sen*	*Spe*	*Pre*
U-Net	0.6254	0.7635	0.9119	0.7868	0.9448	0.7591
U-Net++	0.6283	0.7664	0.9171	0.7479	0.9593	0.8009
Attention-UNet	0.6308	0.7695	0.9109	0.8160	0.9350	0.7535
DoubleU-Net	0.6447	0.7796	0.9171	0.7937	0.9478	0.7795
U2-Net	0.6364	0.7745	0.9107	**0.8347**	0.9308	0.7350
Swin-Unet	0.6252	0.7663	0.9134	0.7675	0.9506	0.7776
SegFormer	0.6392	0.7739	0.9226	0.7307	**0.9702**	**0.8440**
MISSFormer	0.6619	0.7929	0.9243	0.7908	0.9584	0.8105
CACDU-Net	**0.6718**	**0.8001**	**0.9252**	0.8225	0.9483	0.7917

a smaller number of images, resulting in shorter training time and faster convergence of the network. However, the lesions in this dataset are shallower and have blurred edges, making segmentation more difficult.

Table 5 presents the segmentation performance comparison results obtained by experimenting on this dataset (the best result on each metric is shown in **bold**). Again, the proposed CACDU-Net model outperformed all mainstream models based on the two core evaluation metrics, by scoring 0.0099 and 0.0072 points higher than the first runner-up (MISSFormer) for *IoU* and *DSC*, respectively. In addition, based on *accuracy*, CACDU-Net also outperformed all mainstream models by leaving behind the first runner-up (MISSFormer) by 0.0009 points. According to the other three evaluation metrics used, the proposed CACDU-Net model also performed relatively well, by taking correspondingly the second place on *sensitivity*, fourth place on *precision*, and fifth place on *specificity*. Not achieving the best results on these three metrics indicates that the proposed model has certain shortcomings and limitations in accurately detecting diseased regions and excluding non-diseased regions.

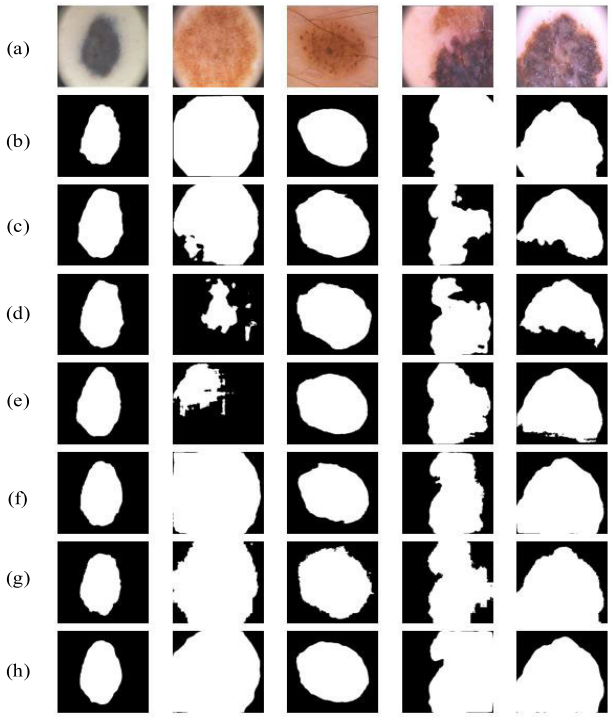


FIGURE 12. Sample segmentations performed on the PH2 dataset: (a) original image; (b) ground truth; (c) U-Net results; (d) U-Net++ results; (e) Attention-UNet results; (f) U2-Net results; (g) Swin-Unet results; (h) CACDU-Net results.

Figure 13 illustrates the loss variation curves of the proposed CACDU-Net model on both the training and validation sets, as well as its DSC and IoU training and validation curves.

Figure 14 shows the ROC curves of the compared models, along with their AUC values, achieved on this dataset. As can be seen from this figure, the proposed CACDU-Net model clearly outperforms all other models, as its ROC curve is the closest one to the upper left corner, which indicates the highest overall accuracy.

The visual comparison of the skin lesion segmentation results achieved by different models on this dataset, shown in Figure 15, illustrates that existing mainstream models can effectively predict larger lesions, while CACDU-Net has a significant advantage in predicting multiple smaller lesions.

4) ABLATION STUDY

In order to verify whether each of the newly designed modules does indeed improve the network performance, ablation study experiments were conducted with U-Net and DoubleU-Net models, used as baselines, on the ISIC2018 dataset. The results of these experiments are shown in Tables 6 and 7 (the best result on each metric is shown in**bold**).

Considering the challenges posed by lesions with different shapes, colors, and blurry edges, contained in the images, the step-by-step addition of designed modules to U-Net and

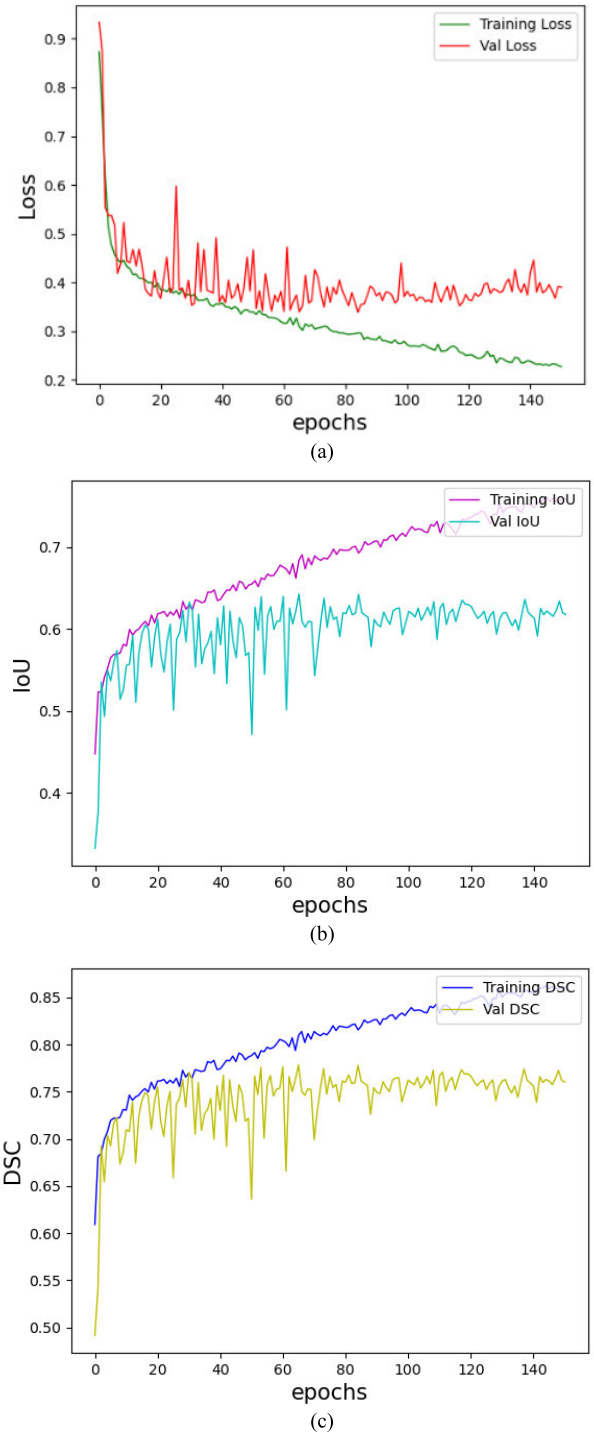


FIGURE 13. Training and validation process of CACDU-Net on the private dataset: (a) training and validation loss curves; (b) IoU training and validation curves; (c) DSC training and validation curves.

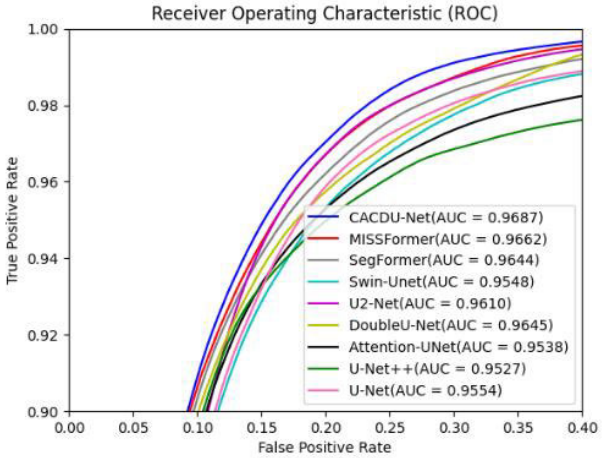


FIGURE 14. ROC curves and AUC values of different models on the private dataset.

FIGURE 15. Sample segmentations performed on the private dataset: (a) original image; (b) ground truth; (c) U-Net results; (d) U-Net++ results; (e) Attention-UNet results; (f) U2-Net results; (g) Swin-Unet results; (h) CACDU-Net results.

DoubleU-Net demonstrated gradual improvement, compared to the baselines and configurations used in the previous steps, on five (out of six) evaluation metrics compared to U-Net and on four evaluation metrics compared to DoubleU-Net (including the main two metrics –*IoU* and *DSC*). For instance, the combined integration of all three designed modules into the original U-Net model gave up the first place (to the 'U-Net+ConvNeXt +ACASPP' configuration) only on *precision* (by only 0.0014 points). For DoubleU-Net, the 'DoubleU-Net+ConvNeXt +ACASPP' configuration demonstrated the best result for *precision*; however, for *sensitivity*, the best result was achieved by the baseline.

On the other hand, these experiments also demonstrated that DoubleU-Net has overall better performance than U-Net. Thus, it was preferred as a basis for the development of the proposed CACDU-Net model.

TABLE 6. Ablation study results using U-Net as a baseline.

Network	*IoU*	*DSC*	*Acc*	*Sen*	*Spe*	*Pre*
U-Net	0.7887	0.8784	0.9508	0.8760	0.9717	0.9182
U-Net+ConvNeXt	0.8116	0.8931	0.9568	0.8743	0.9781	0.9210
U-Net+ConvNeXt +ACASPP	0.8169	0.8969	0.9581	0.8754	0.9800	**0.9261**
U-Net+ConvNeXt +ACASPP+CACB	**0.8222**	**0.9002**	**0.9590**	**0.8833**	**0.9801**	0.9247

TABLE 7. Ablation study results using DoubleU-Net as a baseline.

Network	*IoU*	*DSC*	*Acc*	*Sen*	*Spe*	*Pre*
DoubleU-Net	0.8238	0.9014	0.9572	**0.9271**	0.9643	0.8812
DoubleU-Net +ConvNeXt	0.8313	0.9062	0.9608	0.8985	0.9771	0.9185
DoubleU-Net+ConvNeXt +ACASPP	0.8386	0.9109	0.9639	0.8918	0.9779	**0.9355**
DoubleU-Net +ConvNeXt +ACASPP+CACB (**CACDU-Net**)	**0.8427**	**0.9134**	**0.9641**	0.9065	**0.9790**	0.9252

V. CONCLUSION

Fast and accurate segmentation of skin lesions is crucial for the subsequent treatment of melanoma and other skin cancers. Traditional methods are time-consuming and labor-intensive, heavily dependent on tuning a large number of parameters. In light of this, the paper has proposed a newly designed U-shaped encoder-decoder neural network model, called CACDU-Net. Firstly, it utilizes a pre-trained ConvNeXt-T network as an encoding part to provide rich image features, which allowed it to achieve high values of evaluation metrics at the beginning of the training, thus increasing the network's inference speed. Secondly, the proposed model uses specially designed ConvNeXt Attention Convolutional Blocks (CACB) to provide attention information in both channel and spatial dimensions, focusing on the lesion itself rather than on irrelevant information such as body hair, bubbles, vessels, and measurement scales. Additionally, the use of a stacked U-shaped architecture perfectly combines multi-level features, capturing long-term dependencies in obtaining a global contextual view to help the network achieve accurate segmentation of skin lesions. Thirdly, CACDU-Net utilizes a newly designed ACASPP module, inserted between the encoding and decoding parts, to provide multi-scale semantic information to the network, which is helpful for identifying lesions of different sizes. Based on ASPP, ACASPP adds an asymmetrically structured dilated convolution to finely extract multi-scale information and enhances the network's robustness. In terms of the loss function, a weighted sum of the commonly used binary cross entropy (BCE) and Dice similarity coefficient (DSC) loss functions is used to define a new loss function for solving the problem of extremely uneven numbers of positive and negative samples.

Results, obtained by experiments conducted on three skin lesion image datasets, confirmed that the proposed CACDU-Net model outperforms all existing mainstream models on at least half of the six evaluation metrics used, including the main two metrics for image segmentation evaluation, namely *IoU* and *DSC*. In addition, the proposed model demonstrated robustness and strong adaptability to multi-interference images, at the expense of utilizing a relatively large and computationally expensive neural network.

Importantly, the designed modules proposed in this paper can be used on their own in various U-shaped encoding-decoding networks to enhance their segmentation performance, which constitutes an additional contribution made in the area for use in practical applications.

In the future, we plan to explore the following research routes. Firstly, due to the width of the network (96, 192, 384, 768) in the encoding stage of the ConvNeXt-T structure in Network1, if some appropriate operations can be added to fully combine it with the U-Net network structure, it may further improve the segmentation accuracy. Secondly, we will explore simple post-processing methods, such as connected component analysis, constrained optimization, and linear or nonlinear smoothing, which may also help improve network performance. Thirdly, we will attempt to apply the proposed model to other medical imaging-related tasks, such as lung segmentation, heart segmentation, breast segmentation, and retinal vessel segmentation. We think that using the proposed CACDU-Net model for performing these medical image segmentation tasks, combined with appropriate preprocessing and post-processing techniques, can yield more advanced segmentation results.

REFERENCES

[1] K. T. Flaherty, "Targeting metastatic melanoma," *Annu. Rev. Med.*, vol. 63, no. 1, pp. 171–183, Feb. 2012.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA, A Cancer J. Clinicians*, vol. 69, no. 1, pp. 7–34, Jan. 2019.

[3] A. H. Ali, J. Li, and G. Yang, "Automating the ABCD rule for melanoma detection: A survey," *IEEE Access*, vol. 8, pp. 83333–83346, 2020.

[4] Y. T. Kelman, S. Asraf, N. Ozana, N. Shabairou, and Z. Zalevsky, "Optical tissue probing: Human skin hydration detection by speckle patterns analysis," *Biomed. Opt. Exp.*, vol. 10, no. 9, pp. 4874–4883, 2019.

[5] S. W. Menzies, "A method for the diagnosis of primary cutaneous melanoma using surface microscopy," *Dermatologic Clinics*, vol. 19, no. 2, pp. 299–305, Apr. 2001.

[6] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino, "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis," *J. Amer. Med. Assoc. Dermatol.*, vol. 134, no. 12, pp. 1563–1570, Dec. 1998.

[7] M. A. Kassem, K. M. Hosny, R. Damasevicius, and M. M. Eltoukhy, "Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review," *Diagnostics*, vol. 11, no. 8, p. 1390, Jul. 2021.

[8] R. Baig, M. Bibi, A. Hamid, S. Kausar, and S. Khalid, "Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images—A review," *Current Med. Imag. Formerly Current Med. Imag. Rev.*, vol. 16, no. 5, pp. 513–533, May 2020.

[9] M. Silveira, J. C. Nascimento, J. S. Marques, A. R. S. Marcal, T. Mendonca, S. Yamauchi, J. Maeda, and J. Rozeira, "Comparison of segmentation methods for melanoma diagnosis in dermoscopy images," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 1, pp. 35–45, Feb. 2009.

[10] M. E. Celebi, Y. A. Aslandogan, and P. R. Bergstresser, "Unsupervised border detection of skin lesion images," in *Proc. Int. Conf. Inf. Technol., Coding Comput. (ITCC)*, Las Vegas, NV, USA, 2005, pp. 123–128.

[11] M. E. Celebi, H. A. Kingravi, H. Iyatomi, J. Lee, Y. A. Aslandogan, W. Van Stoecker, R. Moss, J. M. Malters, and A. A. Marghoob, "Fast and accurate border detection in dermoscopy images using statistical region merging," *Proc. SPIE*, vol. 6512, pp. 1297–1306, Mar. 2007.

[12] M. E. Celebi, Q. Wen, S. Hwang, H. Iyatomi, and G. Schaefer, "Lesion border detection in dermoscopy images using ensembles of thresholding methods," *Skin Res. Technol.*, vol. 19, no. 1, pp. 252–258, Feb. 2013.

[13] R. Garnavi, M. Aldeen, M. E. Celebi, G. Varigos, and S. Finch, "Border detection in dermoscopy images using hybrid thresholding on optimized color channels," *Computerized Med. Imag. Graph.*, vol. 35, no. 2, pp. 105–115, Mar. 2011.

[14] P. Schmid, "Segmentation of digitized dermatoscopic images by two-dimensional color clustering," *IEEE Trans. Med. Imag.*, vol. 18, no. 2, pp. 164–171, 1999.

[15] A. Agarwal, A. Issac, M. K. Dutta, K. Riha, and V. Uher, "Automated skin lesion segmentation using K-means clustering from digital dermoscopic images," in *Proc. 40th Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2017, pp. 743–748.

[16] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: A review," *J. Med. Syst.*, vol. 42, no. 11, pp. 1–13, Nov. 2018.

[17] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. W. M. van Uder, F. E. de Leeuw, E. Marchiori, B. van Ginneken, and B. Platel, "Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 1414–1417.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[20] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 2017.

[21] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, and B. Glocker, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Munich, Germany: Springer, 2015, pp. 234–241.

[23] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, and H. Kittler, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," 2019, *arXiv:1902.03368*.

[24] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH^2—A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Osaka, Japan, Jul. 2013, pp. 5437–5440.

[25] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564.

[26] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 11966–11976.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[28] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Granada, Spain: Springer, 2018, pp. 3–11.

[29] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.

[30] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *J. Amer. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 544–551, 2011.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and J. Uszkoreit, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[32] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-UNet: UNet-like pure transformer for medical image segmentation," in *Computer Vision—ECCV*. Tel Aviv, Israel: Springer, 2023, pp. 205–218.

[33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 9992–10002.

[34] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.

[35] X. Huang, Z. Deng, D. Li, X. Yuan, and Y. Fu, "MISSFormer: An effective transformer for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1484–1494, May 2023.

[36] C. Yang and F. Gao, "EDA-Net: Dense aggregation of deep and shallow information achieves quantitative photoacoustic blood oxygenation imaging deep in human breast," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Shenzhen, China: Springer, 2019, pp. 246–254.

[37] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 1911–1920.

[38] R. Li, C. Duan, and S. Zheng, "MACU-Net: Semantic segmentation from high-resolution remote sensing images," 2020, *arXiv:2007.13083*.

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826.

[41] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[43] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 3–19.

[44] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[45] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[48] Q. Zhou, T. He, and Y. Zou, "Superpixel-oriented label distribution learning for skin lesion segmentation," *Diagnostics*, vol. 12, no. 4, p. 938, Apr. 2022.

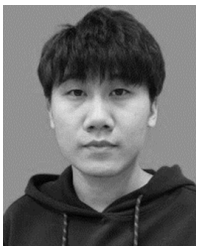
[49] J. Ruan, S. Xiang, M. Xie, T. Liu, and Y. Fu, "MALUNet: A multi-attention and light-weight UNet for skin lesion segmentation," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Las Vegas, NV, USA, Dec. 2022, pp. 1150–1156.

[50] D. Dai, C. Dong, S. Xu, Q. Yan, Z. Li, C. Zhang, and N. Luo, "Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102293.

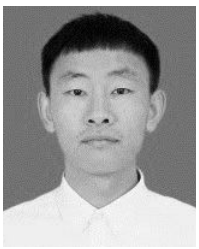
[51] W. Cao, G. Yuan, Q. Liu, C. Peng, J. Xie, X. Yang, X. Ni, and J. Zheng, "ICL-Net: Global and local inter-pixel correlations learning network for skin lesion segmentation," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 1, pp. 145–156, Jan. 2023.

[52] R. Azad, Y. Jia, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "Enhancing medical image segmentation with TransCeption: A multi-scale feature fusion approach," 2023, *arXiv:2301.10847*.
[53] J. Mu, Y. Lin, X. Meng, J. Fan, D. Ai, D. Chen, H. Qiu, J. Yang, and Y. Gu, "M-CSAFN: Multi-color space adaptive fusion network for automated port-wine stains segmentation," *IEEE J. Biomed. Health Informat.*, early access, Feb. 22, 2023.
[54] Z. Song, W. Luo, and Q. Shi, "Res-CDD-Net: A network with multi-scale attention and optimized decoding path for skin lesion segmentation," *Electronics*, vol. 11, no. 17, p. 2672, 2022.

SHENGNAN HAO received the B.S. degree from the North China University of Science and Technology, China, in 1996, and the M.S. degree from the Beijing University of Technology, China, in 2009. She joined the North China University of Science and Technology, in 1996, and became an Associate Professor, in 2009. Her current research interests include complex systems, impulsive systems, and stochastic control.



HAOTIAN WU was born in 1996. He received the B.S. degree from the Jilin Institute of Physical Education, in 2018. He is currently pursuing the master's degree with the North China University of Science and Technology. His research interests include machine vision and graphic image processing.



CHENGYUAN DU was born in 1997. He received the B.S. degree from the North China University of Science and Technology, in 2020, where he is currently pursuing the master's degree. His research interests include recommendation algorithms and natural language processing.

XINYI ZENG was born in 1999. She received the bachelor's degree from the Yancheng Institute of Technology, in 2021. She is currently pursuing the master's degree with the North China University of Science and Technology. Her research interests include machine vision and intelligent medical image processing.

ZHANLIN JI (Member, IEEE) received the M.Eng. degree from Dublin City University, in 2006, and the Ph.D. degree from the University of Limerick, Ireland, in 2010. He is currently a Professor with the North China University of Science and Technology, China, and an Associate Researcher with the Telecommunications Research Centre (TRC), University of Limerick. He has authored/coauthored more than 100 research papers in refereed journals and conferences. His research interests include ubiquitous consumer wireless world (UCWW), the Internet of Things (IoT), cloud computing, big data management, and data mining.

XUEJI ZHANG is currently the Vice President of Shenzhen University and a Professor with the School of Biomedical Engineering, China. His research interests include disciplines of chemistry, biology, materials, and medicine, with an emphasis on studies of biosensing, biomedicine, and biomaterials. He has received numerous national and international awards and honors, including a member of the Russian Academy of Engineering; a fellow of the American Institute for Medical and Bioengineering; a fellow of the Royal Chemical Society; a National Innovation Award in China; a Scientist of the Year in China, and a Simon Fellow of the ICSC-World Laboratory. He serves as the Co-Editor-in-Chief for *Sensors & Diagnostics* and has been an editorial member of 24 international journals.

IVAN GANCHEV (Senior Member, IEEE) received the engineering and Ph.D. degrees (summa cum laude) from the Saint-Petersburg University of Telecommunications, in 1989 and 1995, respectively. He is currently with the University of Limerick, Ireland; the University of Plovdiv "Paisii Hilendarski;" and the IMI-BAS, Bulgaria. He participated in more than 40 international and national research projects. He has authored/coauthored one monographic book, three textbooks, four edited books, and more than 300 research papers in refereed international journals, books, and conference proceedings. He has served on the TPC for more than 390 prestigious international conferences/symposia/workshops. He is on the editorial board and has served as the guest editor for multiple international journals. He is also an International Telecommunications Union (ITU-T) Invited Expert and an Institution of Engineering and Technology (IET) Invited Lecturer.